# АВТОМАТИЗАЦИЯ АНАЛИЗА МНЕНИЙ ПОЛЬЗОВАТЕЛЕЙ ОБ УСЛУГАХ В СФЕРЕ ТУРИЗМА НА ОСНОВЕ ОТЗЫВОВ В СОЦИАЛЬНЫХ СЕТЯХ

**Е. А. Бурдук**

*Учреждение образования «Гомельский государственный технический университет имени П. О. Сухого», Республика Беларусь*

Научный руководитель И. А. Мурашко

В настоящее время количество публикуемых отзывов достигает нескольких десятков тысяч и сбор информации в интернете вручную становится трудоемкой, рутинной и отнимающей много времени работой. Также активное развитие социальных сетей, форумов и блогов увеличивает интерес к задаче автоматизированного анализа мнений пользователей сети интернет по различным вопросам. В связи с этим возникает потребность в автоматическом сборе информации и автоматической оценке текста [1].

Целью этой работы является написание парсера, позволяющего осуществлять поиск отзывов об отелях, турфирмах и странах мира. Информация, полученная после работы парсера, обрабатывается и проводится анализ тональности текста, который дает возможность получить представление об эмоциональном отношении авторов к объектам, о которых идет речь в тексте.

Парсинг представляет собой обработку информации, расположенной на страницах сайтов и выделение из нее необходимых данных. Процесс парсинга выполняется

специальной программой-парсером. Программа-парсер быстро изучит большое количество сайтов, аккуратно отделит нужную информацию от программного кода и безошибочно выберет нужную информацию.

Программа парсинга позволяет пользователям сети интернет отказаться от ручной работы, отнимающей много времени на поиск отзывов о некотором объекте и предоставить требуемую информацию пользователям. Парсер предоставляет информацию в определенном виде, который также задается разработчиком программы.

Весь процесс парсинга можно разделить на несколько этапов:

1) получение исходного кода интернет-страницы;

2) проведение анализа полученных данных путем извлечения требуемой информации из кода разметки;

3) обработка и преобразование данных в необходимый формат для дальнейшего использования;

4) генерация результата и его вывод в файл или на экран – завершающий этап парсинга.

Для реализации парсинга был выбрана библиотека *AngleSharp*, которая представляет собой быстрый парсер с удобным *API*. *API* построен на базе официальной спецификации по *JavaScript HTML DOM*, которая относится к *W3C* стандарту, поддерживаемого всеми современными браузерами. *DOM* описывает структуру веб-страницы в виде древовидного представления и предоставляет возможность получить доступ к отдельным элементам веб-страницы.

Для осуществления задачи эмоциональной оценки текста был использован подход на основе машинного обучения с учителем [2]. Данный подход является одним из наиболее распространенных подходов, применяемых в исследованиях. В основе этого подхода лежит обучение машинного классификатора на предварительно собранной коллекции текстов, каждому из которых заранее указывается правильный тип тональности. После чего полученная модель используется для анализа новых документов.

Для осуществления алгоритма классификации был выбран наивный байесовский классификатор, являющийся одним из самых простых в тестировании [3]. В основе наивного байесовского классификатора лежит теорема Байеса, которая позволяет определить вероятность события при условии, что произошло другое взаимозависимое событие:

$$P(c \mid d) = \frac{P(d \mid c) \cdot P(c)}{P(d)}, \tag{1}$$

где $P(c \mid d)$ – вероятность, что документ $d$ принадлежит классу $c$; $P(d \mid c)$ – вероятность встретить документ $d$ среди всех документов класса $c$; $P(c)$ – безусловная вероятность встретить документ класса $c$ в обучающей выборке документов; $P(d)$ – безусловная вероятность встретить документ класса $d$ в обучающей выборке документов.

Цель классификации состоит в том, чтобы понять, к какому классу принадлежит документ, поэтому нам нужна не сама вероятность, а наиболее вероятный класс. Байесовский классификатор использует оценку апостериорного максимума (*Maximum a posteriori estimation*) для определения наиболее вероятного класса. То есть необходимо рассчитать вероятности, с которыми документ принадлежит к каждому из классов и выбрать класс, обладающий максимальной вероятностью.

Данная оценка определяется по формуле (2):

$$C_{map} = \arg\max_{c \in C} = \frac{P(d \mid c) \cdot P(c)}{P(d)}, \tag{2}$$

где $C_{map}$ – оценка апостериорного максимума.

Необходимо рассчитать вероятность для всех классов и выбрать тот класс, который обладает максимальной вероятностью.

Байесовский классификатор представляет документ как набор слов вероятности, которые условно не зависят друг от друга. Исходя из этого предположения, условная вероятность документа аппроксимируется произведением условных вероятностей всех слов, входящих в документ:

$$P(d \mid c) \approx P(\omega_1 \mid c) P(\omega_2 \mid c) \ldots P(\omega_n \mid c) = \prod_{i=1}^{n} P(\omega_1 \mid c), \tag{3}$$

где $P(\omega_i \mid c)$ – условные вероятности всех слов входящих в документ.

Оценка вероятностей и $P(c)$ и $P(\omega_i \mid c)$ осуществляется на обучающей выборке. Вероятность класса можно оценить по формуле (4):

$$P(c) = \frac{D_c}{D}, \tag{4}$$

где $D_c$ – количество документов, принадлежащих классу $c$; $D$ – общее количество документов в обучающей выборке.

Оценка вероятности слова в классе определяется по формуле (5):

$$P(\omega_i \mid c) = \frac{W_{ic}}{\sum_{i' \in V} W_{i'c}}, \tag{5}$$

где $W_{ic}$ – описывает, сколько раз слово встречается в файлах класса $c$; $W_{i'c}$ – количество слов во всех документах класса $c$.

Если на этапе классификации встречается слово, которого нет в обучающей выборке, то значения, а следственно $W_{ic}$ и $P(\omega_i \mid c)$, будут равны нулю. Это приведет к тому, что документ с этим словом нельзя будет классифицировать, так как он будет иметь нулевую вероятность по всем классам.

Типичным решением проблемы неизвестных слов является аддитивное сглаживание (сглаживание Лапласа). Идея заключается в том, что необходимо прибавить единицу к частоте каждого слова:

$$P(\omega_i \mid c) = \frac{W_{ic} + 1}{\sum_{i' \in V} (W_{i'c} + 1)} = \frac{W_{ic} + 1}{|V| + \sum_{i' \in V} W_{i'c}}. \tag{6}$$

Логически данный подход смещает оценку вероятностей в сторону менее вероятных исходов. Таким образом, слова, которых нет на этапе обучения модели, получают пусть маленькую, но не нулевую вероятность.

Разработанное приложение позволяет пользователю получить информацию об отелях, турфирмах и турах на основе мнений, составленных другими пользователями сети интернет.

Результат работы приложения был протестирован на отзывах, взятых с сети интернет об услугах в сфере туризма. Каждый отзыв был оценен по шкале от –10 до 10 и пользователю был представлен результат оценки. Приложение позволяет пользователям получать требуемую информацию быстро, не затрачивая на это свое время.

Литература

1. Pang, B. Opinion Mining and Sentiment Analysis / B. Pang, L. Lee. – Philadelphia: Now Publishers Inc, 2008. – P. 35–80.
2. Heerschop, B. Polarity analysis of texts using discourse structure / B. Heerschop, F. Goossen, A. Hogenboom // Proceedings of the 20th ACM international conference on Information and knowledge management. – 2011. – P. 1061–1070.
3. Осокин, В. В. Анализ тональности русскоязычного текста / В. В. Осокин, М. В. Шегай // Интеллектуал. системы. Теория и приложения. – 2014. – № 3. – С. 163–174.